Final Project
CUNY621 – MS in Data Science
Group5, April 21, 2019

PRESENCE OF HEART DISEASE PREDICTION

1. DESCRIPTION

With nearly 18MM deaths in 2015[1] , cardiovascular diseases (CVD) are the leading cause of death globally and growing in the developing world[2]. CVD is a disease class which includes heart attacks, strokes, heart failure, coronary artery disease, arrythmia, venous thrombosis, and other conditions. About half of all Americans (47%)[3] have at least one of three key risk factors for heart disease: high blood pressure, high cholesterol, and smoking.

CVD's high mortality rate is particularly poignant as researchers estimate that up to 90% of heart disease deaths[4] could be prevented. Typical means of detection include electrocardiograms (ECGs), stress tests, and cardiac angiograms, all of which are expensive. Risk evaluation screenings require blood samples, which are assessed alongside risk factors like tobacco use, diet, sleep disorders, physical inactivity, air pollution, and others.

There is a need for more efficient, scalable (i.e. cost-effective even in the developing world), and – optimally – non-invasive means of early detection that can trigger medical interventions, prompt preventive care by physicians, and/or engender behavioral change on the part of those prone to or suffering from CVD.

Applying data mining techniques to CVD datasets to predict risk based on existing or easy-to-collect health data could improve healthcare outcomes and mortality rates. Ongoing experimentation has been focused on several datasets[5] (below):

Database Donor[6]	**Observations**
Cleveland Clinic Foundation	303
Hungarian Institute of Cardiology	294
University Hospital, Zurich	123
Long Beach V.A. Medical Center	200

All of these medical datasets include a relatively small number of observations, which is not uncommon given the costs of experimental data and privacy risk of observational data. Additionally, the attributes recorded are not consistent between these datasets, making meaningful comparison difficult, and the Hungarian, Swiss, and Long Beach sets are missing many variables. Accordingly, as the most complete of the group, the Cleveland dataset is the most frequently used in data science experimentation.

2. DATA

The Cleveland dataset contains 303 observations and 76 attributes in total. However, all published experiments refer to a subset of 14 of these attributes, which are available via the UCI machine learning database[7].

The dataset includes 13 features and 1 target variable. Data science experiments have concentrated on distinguishing the presence of heart disease (with values of 1 through 4) from its absence (a value of 0).

Feature	Variable	Description	Type
Age	age	In years	Continuous
Sex	sex	Gender	Categorial
Chest pain type	cp	Scale of 0 to 4 (typical angina, atypical angina, non-angina pain, asymptomatic)	Categorial
Resting blood pressure	trestbps	Diastolic blood pressure in mmHg	Continuous
Cholesterol	chol	Serum cholesterol (mg/dl)	Continuous
Fasting blood sugar	fbs	Greater than 120mg/dl, value of 0 or 1	Categorical
Resting ECG	restecg	Value of 0, 1, or 2	Categorical
Maximum heartrate achieved	thalach	Maximum heartrate from thallium test[8]	Continuous
Exercise-induced angina	exang	Value of 0 or 1	Categorical
Old-peak	oldpeak	ST depression induced by exercise relative to rest [9]	Continuous
Slope-peak	slope	Slope of peak exercise ST segment, value of 1, 2, or 3	Categorical
Coronary artery disease	ca	Number of major vessels (0-3) colored by fluoroscopy	Categorical
Exercise thallium	thal	Exercise thallium scintigraphic defects, vales of 3 (normal), 6 (fixed defect), or 7 (reversible defect)	Categorial

3. LITERATURE REVIEW

For the purpose of this project proposal we conducted a high-level review of the existing literature. Shouman et al.[10] presents an exhaustive review work between 2000 and 2016, detailing a wide range of classification techniques used on the above CVD datasets, including Logistic Regression, Decision Trees, Random Forests, KNN, Naïve Bayes, a variety of Neural Network approaches, Multilayer Perceptron, Support Vector Machines, Associative Classifiers, and others. A smaller but equally diverse range have been applied specifically to the Cleveland dataset, yielding diagnostic accuracy peaking in the mid- to high 80% range. This audit concludes that performance is strongest when more than one technique is integrated, though results vary across the literature due to discretization methods and other differences.

Based on this high-level review, we focused on a few pieces of research in particular:

- The aim of Shouman et al.'s work is to evaluate a potential low-cost heart disease expert system risk evaluation tool leveraging non-invasive data attributes. This is explored by evaluating the Cleveland dataset alongside another dataset from Canberra not available via the UCI machine learning repository. When constrained to Cleveland's non-invasive data attributes, the best performance is seen with a combination of age, sex, and resting blood pressure. This line of research also explores integrating K-means clustering with decision tree models to improve accuracy.
- Assari et al.'s[11] broader data-mining focus finds that SVM and Naïve Bayes outperform KNN (of K=7) and Decision Tree in terms of accuracy when using 10-fold cross-validation. Its results identify the most important features as chest pain type, exercise thallium, and coronary artery disease.
- Sabay et al.[12] seek to assess the application of ML techniques requiring more observations to the Cleveland dataset so improve its generalizability. To that end, a surrogate synthetic dataset is bootstrapped using the Synthpop package in R. Logistic Regression is found to be more accurate and stable than Random Forest and Decision Tree methods, both for the original dataset as well as a 50,000-observation surrogate. An ANN perceptron model built on a 60,000-observation surrogate dataset achieves accuracy and recall above 95%.

4. HYPOTHESIS AND DISCUSSION

A multitude of approaches and methodologies have been undertaken in an attempt to predict the presence of heart disease using the 14 variables most commonly selected from the Cleveland dataset. Sabay et al. saw the strongest results with a logistic model used on surrogate datasets; Shouman et al. with a combination of KNN + Decision Trees; and Assari et al. with SVM, Naïve Bayes, and cross-validation.

For this project, we intend to combine elements of these different approaches in ways not encountered in our literature review. We will evaluate the performance of interpretable models – logistic classification and decision tree, and perhaps others – using bootstrapped datasets and cross-validation to explore the implications for classification accuracy. While some of these techniques will be new to the group, we think the exercise will serve as valuable learning experience and exposure to data mining techniques.

[1] Mendis, S., Puska, P., & Norrving, B. (Eds.). (2011). *Global atlas on cardiovascular disease prevention and control*. Geneva: World Health Organization. Retrieved April 20, 2019, from http://whqlibdoc.who.int/publications/2011/9789241564373_eng.pdf?ua=1.

[2] GBD 2015 Mortality and causes of death collaborators (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: A systematic analysis for the Global Burden of Disease Study 2015. *Lancet (London, England)*, *388*(10053), 1459–1544. doi:10.1016/S0140-6736(16)31012-1

[3] Heart disease risk factors. (n.d.). Retrieved April 20, 2019, from https://www.cdc.gov/heartdisease/risk_factors.htm

[4] O'Donnell, M. J., Chin, S. L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., … Yusuf, S. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): A case-control study. *The Lancet*, 388(10046), 761-775. doi:10.1016/s0140-6736(16)30506-2

[5] (n.d.). Retrieved April 20, 2019, from http://archive.ics.uci.edu/ml/datasets/heart+disease

[6] Aha, D. (1988). *heart-disease.names*. Retrieved from https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names

The authors of the databases have requested:

> ...that any publications resulting from the use of the data include the names of the principal investigator responsible for the data collection at each institution. They would be:
>
> 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
> 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
> 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
> 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D

[7] (n.d.). Retrieved April 20, 2019, from http://archive.ics.uci.edu/ml/datasets/heart+disease

[8] Thallium stress test: Purpose, procedure, and risks. (n.d.). Retrieved April 20, 2019, from https://www.healthline.com/health/thallium-stress-test

[9] ST depression. (2018, October 29). Retrieved April 20, 2019, from https://en.wikipedia.org/wiki/ST_depression

[10] Shouman, M. M. (2014, March). *Prototype development of a novel heart disease risk evaluation tool using data mining analysis*. Retrieved from http://unsworks.unsw.edu.au/fapi/datastream/unsworks:12635/SOURCE02?view=true

[11] Assari, R., Azimi, P., & Reza Taghva, M. (2017). Heart disease diagnosis using data mining techniques. *International Journal of Economics & Management Sciences*, *06*(03). doi:10.4172/2162-6359.1000415. Retrieved from https://www.omicsonline.org/open-access/heart-disease-diagnosis-using-data-mining-techniques-2162-6359-1000415.pdf

[12] Jaceldo-Siegl, K. (n.d.). Overcoming small data limitations in heart disease prediction by using surrogate data. Retrieved from https://scholar.smu.edu/datasciencereview/vol1/iss3/12/